What's New

IMG 2.5 is the **13th** release of the Integrated Microbial Genomes (IMG) genomic data management and analysis system. **IMG 2.5** was released on **March 31st, 2008.**

IMG 2.5 Content

Genomes

The content of **IMG 2.5** has been updated with new microbial genomes available in **RefSeq version 27** (January 6, 2008).

IMG 2.5 contains a total of **3,818** genomes consisting of **865** bacterial, **53** archaeal, **40** eukaryotic genomes, **2,123** viruses (including bacterial phages), and **737** plasmids that did not come from a specific microbial genome sequencing project. Among these genomes:

- **3,503** are finished genomes, and **315** are draft genomes.
- **272** are **JGI** sequenced genomes: **206** finished and **66** draft genomes. JGI genomes are also available through individual microbial portals at http://genome.jgi-psf.org/microbial.

Note that **20** microbial genomes from **IMG 2.4** were **replaced** in **IMG 2.5** because (1) a "Draft" genome has been replaced by its "Finished" version or (2) the composition of the genome has changed through the addition of new replicons, that is, plasmids or chromosomes. For replaced genomes, whenever possible, the gene object identifiers (gene OIDs) for the protein-coding genes (CDS) were mapped to their new version in IMG. 2.5. See IMG Data Evolution History for details.

Plasmid names were curated for **27** plasmids by adding strain names to organism name when available from publications or other sources.

Annotations

IMG Terms and Pathways

IMG's native controlled vocabularies for gene function (IMG Terms) and organism-independent functional hierarchies (IMG Pathways and PartsLists) have been extended. IMG 2.5 has **3,245** IMG Terms, and **559** IMG Pathways and Parts Lists. In addition, **574,148** IMG genes are associated with IMG Terms.

rRNAs and tRNAs

tRNA and rRNA genes (23S, 16S and 5S) that are missing from the original RefSeq genome files are added using tRNAscan-SE v1.23 for tRNA genes and similarity comparisons to existing RNA genes. In IMG 2.5 **3,762** tRNA and **1,445** tRNA genes were added in **142** genomes.

Chromosomal Cassettes

A ***chromosomal cassette*** is defined as a stretch of protein coding genes with intergenic distance smaller or equal to 300 base pairs. The genes must be on the same strand or divergent; convergent genes are not allowed to participate in the formation of a chromosomal cassette. Groups of at least two common genes between two or more chromosomal cassettes are defined as ***conserved chromosomal cassettes***. In order to identify common genes between chromosomal cassettes, genes need to be assigned to groups of equivalent genes. For this grouping, the commonly accepted clusters of orthologous genes (COG), Pfam assignments, and clusters of bidirectional best hits implemented in IMG using the MCL algorithm were used. If a protein consists of multiple clusters, such as in a gene fusion or multiple Pfam domains, each individual domain is included in the chromosomal cassette.

IMG Ortholog Clusters

IMG ortholog clusters are formed using the Markov Cluster Algorithm (MCL) applied on bidirectional best hits between proteins. A *conservation score* is calculated to normalize the strength of similarity. This score consists of the bit score between two sequences divided by bit score of the sequences when BLASTed against itself (self bit score). More precisely, it is

$$cons\_score_{xy} = bit\_score_{xy} / max(bit\_score_{xx}, bit\_score_{yy})$$

where *x* and *y* are two separate sequences. The *mcl* tool is run with default parameters.

IMG 2.5 User Interface

The User Interface has been extended in order to improve its overall usability.

Genomes

Genome Details

In the **Genome Statistics** table, each **list of genes** associated with a specific **functional** category (e.g., Pfam, Enzymes, IMG terms) leads to a table listing the gene sub-lists associated with individual functions. Each column in this table can be sorted.

Genes

Gene Details

Gene Information Section

- **GO terms**: only molecular functions are displayed.
- An **IMG Clusters** sub-section has been added under **Pathway Information:**
    - **Chromosomal Cassette** lists the cassette (see definition above) containing the current gene. **Details** regarding the cassette can be displayed with genes labeled by their COG, Pfam, or IMG ortholog cluster association.
    - **Protein Cluster Context Analysis** allows examining the functional correlations of the current gene based on its COG, Pfam or IMG ortholog cluster association. **Context Analysis** starts with the current's gene (so called "query") COG, Pfam or IMG ortholog cluster. For a query protein cluster ***P***, the context analysis page displays:
        - A **summary** with the number of: (i) genomes containing genes associated with ***P***, (ii) other protein clusters (of the same type as ***P***) within the same chromosomal cassettes with ***P***, (iii) other protein clusters (of the same type as ***P***) within the same conserved chromosomal cassettes with ***P***, (iv) number of other protein clusters (of the same type as ***P***) fused with ***P.***
        - A **table** containing **connectivity ratios** between ***P*** and protein clusters that are either fused with ***P*** or appear with ***P*** within the same chromosomal cassette o conserved chromosomal cassette. This table also contains the data used for computing the ratios.
    - **IMG Ortholog Cluster** lists the IMG ortholog cluster (see definition above) containing the current gene.

Evidence for Function Prediction Section

- **Chromosome Viewer:** intergenic regions can be moused over for viewing details.

- A new **Conserved Neighborhood** sub-section has bee added after the **Neighborhood** sub-section of the **Evidence for Function Prediction** section**.** This sub-section includes the **Ortholog Neighborhood Viewer** and a new **Chromosomal Cassette Viewer.** Chromosome cassettes can be viewed with genes labeled by their COG, Pfam, or IMG ortholog cluster association.

Homolog Display Section

Homolog **s**elections open a **new page** listing the results.

Analysis Carts

Gene Cart

ClustalW alignment result includes the list of genes that are aligned with links to their respective Gene Details pages.

Function Cart

- Individual **COG, Pfam, Enzyme, TIGRfam** functions are linked to their native definition or detail pages**.**
- **Function Profile:** the page displaying the result of a Function Profile includes a "Show all Genes" link that leads to the list of all the genes that are associated with functions involved in this profile.